

Using Apollo at the i5k Workspace@NAL

NAL USDA-ARS

<https://i5k.nal.usda.gov>

April 24th, 2018



Agenda

- Manual annotation general overview
- 15k Workspace tools for manual annotation
 - BLAST, Clustal, HMMER
 - Apollo
- Manual annotation example: preparation
- Manual annotation live example

Other resources

- Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
 - <https://www.slideshare.net/MonicaMunozTorres/presentations>
 - I recommend these slides if you need more background:
 - <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>
 - Note - there are two versions of Apollo. The i5k Workspace still uses the older version with a slightly different interface
 - If you are new to Apollo, or need a refresher, we **highly recommend** that you review one of her presentations
- The official Apollo annotation guide:
 - <http://genomearchitect.org/users-guide/>
- Other manual curation tutorials:
 - <https://i5k.nal.usda.gov/manual-curation-example>
 - <http://genomecuration.github.io/genometrain/d-feature-curation-crossing/>

Manual annotation general overview

What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Often, but not always: drawing on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model
 - Structural annotation – defining the gene structure (e.g. exon boundaries)
 - Functional annotation – describing the gene function (e.g. its name)

Why manually annotate?

- “Incorrect annotations poison every experiment that makes use of them”
- “Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes.”
 - Yandell and Ence 2012, doi:10.1038/nrg3174

General process of manual annotation

1. Select a chromosomal region of interest (e.g. scaffold)
 1. E.g. find sequence of interest from one or several other species, and align against proteins or genome sequence from your species
2. Select appropriate evidence (tracks in Apollo, or your own files)
3. Determine whether a feature in your evidence provides a reasonable starting gene model
 1. If yes: select and drag the feature to the 'user-created annotations' area, creating an initial gene model. If necessary use editing functions to adjust the model.
 2. If not – get in touch with us!
4. Edit model if necessary
5. Check your edited gene model for integrity and accuracy by comparing it with available homologs
 1. Verify that the gene model is the best representation of the underlying biology
6. Repeat steps 1 through 5 as needed to refine model
7. Add annotation details in the “Information Editor”
 1. Name, symbol, other comments

Adapted from <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>

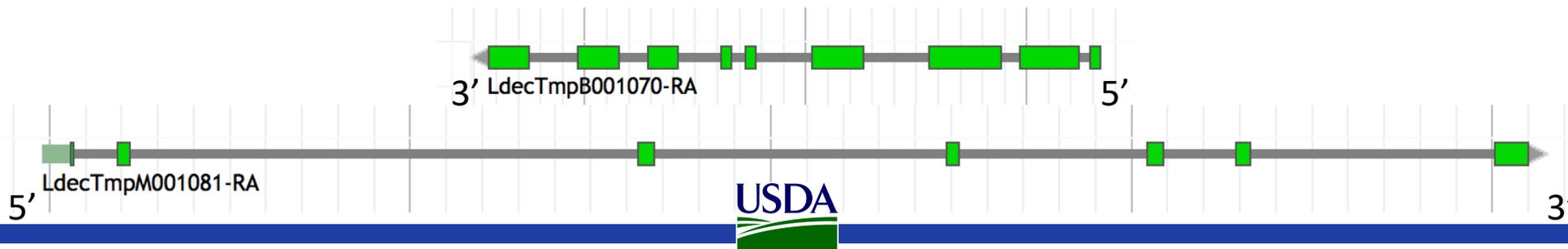
I5k Workspace ‘Etiquette’

1. Use Apollo to improve a gene model in an i5k Workspace assembly.
 1. If you just want to practice – use one of our training instances.
 1. <https://i5k.nal.usda.gov/jbrowseapollo-training>
 2. If you just want to view the data – you probably can get what you want without using Apollo. All of the data that we host is public.
2. Your annotation work is a community effort.
 1. If you notice that someone else is working on your model of choice, get in touch with them (or us) and collaborate – don’t make a 2nd model or delete the other model.
 2. Keep in mind that your work will be used by the scientific community once you’re done.
3. If you publish any of your work generated in the i5k workspace:
 1. Get in touch with the genome contact first (you can find the contact info on the organism page; <https://i5k.nal.usda.gov/species>);
 2. Please cite the i5k Workspace paper! This helps us continue to exist.
 1. <https://doi.org/10.1093/nar/gku983>

Manual annotation: i5k Workspace tools

First, some conventions

- HSP – High scoring pair in BLAST/BLAT alignments
 - The ‘Hits’ in an alignment result set
 - A subsection of a pair of sequences with sufficient score
 - HSPs can change based on the alignment parameters
- Five prime end and three prime end
 - Based on direction of transcription
 - Initiation site is at the five prime end
 - Stop codon is at the three prime end
- In the genome browser, arrowheads indicate direction



i5k Workspace BLAST: one way to access Apollo

The screenshot shows the BLAST interface with the following elements and annotations:

- BLAST Databases:**
 - Organisms:** A list of organisms with *Eurytemora affinis* selected. Annotation: "Select organism" with an arrow pointing to the list.
 - Nucleotide:** "Genome Assembly - Eaff_11172013.genome_new_ids.fa" is selected. Annotation: "Select organism-specific database" with an arrow pointing to this option.
 - Peptide:** "Protein - EAFF_new_ids.faa" is unselected.
- Query Sequence:** A text box contains a peptide sequence: `>FBpp0070332
MDNCDQDASFRLSHIKEEVKPDISQLNDSNN
SSFSPKAESPVPFMQAMSMVHVLPGSNSASS
NNSAGDAQMAQAPNSAG
GSAAAQVQYPPNHPLSGSKHLCSICGDRA
SGKHYGVYCEGCKGFFKTRVRKDLTYACRE`. Annotation: "Paste or upload query sequence(s)" with an arrow pointing to the text box.
- Program:** "tblastn" is selected. Annotation: "Program is automatically selected" with an arrow pointing to the "tblastn" radio button.

URL: <https://i5k.nal.usda.gov/webapp/blast/>

i5k Workspace BLAST: one way to access Apollo

blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qstart
euraff	FBpp0070332	Scaffold427	36.36	77	49	0	419
euraff	FBpp0070332	Scaffold427	26.67	165	83	4	262
euraff	Eaff_11172013.genome_new_ids.fa		59.21	76	31	0	103
euraff	FBpp0070332	Scaffold229	56.52	92	37	1	98
euraff	FBpp0070332	Scaffold200	57.14	91	36	1	99
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold3	85.71	35	5	0	91
euraff	FBpp0070332	Scaffold200	50.62	81	38	1	101

BLAST result page with 4 panels

Click on blue blastdb icon next to your favorite HSP

Blast results are displayed in Apollo

HMMER and Clustal

- Use HMMER to detect remote protein homologs
- <https://i5k.nal.usda.gov/webapp/hmmer/>
- Use Clustal to perform multiple sequence alignments
- <https://i5k.nal.usda.gov/webapp/clustal/>

Tips and Tricks

- The i5k Workspace BLAST results persist for one week
 - You can bookmark and share searches
 - BLAST HSPs are ‘draggable’ and can be used in annotations
- Jbrowse/Apollo URLs can be shared
 - Allow you to share the exact view (including active tracks) with others
 - Great for troubleshooting with collaborators
- In Apollo “walk” feature boundaries
 - Square brackets walk exon boundaries: [and]
 - Curly brackets walk gene boundaries: { and }
- In Apollo, you can pin tracks to the top
- If you know the name or ID of the gene that you’d like to annotate, you can paste it into the search box in Apollo to navigate to it

Manual annotation example: preparation

Annotation Example

- Phosphoenolpyruvate carboxykinase (pepck) in the copepod *Eurytemora affinis*
- Pepck catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).
- More information about the copepod:
https://i5k.nal.usda.gov/Eurytemora_affinis
- Apollo URL:
<https://apollo.nal.usda.gov/euraff/jbrowse/>
 - Note: There are no demo accounts for this species

Notes on *E. affinis* genome/browser

- Big advantage for annotation: lots of RNA-Seq and transcriptome data are available to use as contributing evidence for your gene models
 - Includes strand-specific RNA-Seq
- Disadvantage: No close reference genomes, so it may be harder to find homologs for your genes of interest to inform your annotations.

Available tracks for *E. affinis*

The screenshot shows the Apollo genome browser interface. On the left, the 'Available Tracks' panel is expanded to show the following categories and their counts:

- 0. Reference Assembly: 2
- BCM_v0.5.3: 47
- 1. Gene Sets: 3
 - Primary Gene Sets: Protein Coding: 1
 - EAFF_v0.5.3-Models
 - Supplementary Gene Predictions: 2
 - augustus_masked
 - snap_masked
- 2. Evidence: 2
- 3. Mapped Proteins: 41
- 4. Transcriptome: 1

The 'Transcriptome' category is expanded to show 26 tracks:

- Assembly: 2
 - 075_zz91_transcriptome
 - Mixture of males and females: cufflinks_IGS_UMA1
- Coverage Plots (BigWig): 10
- Mapped Reads: 7
 - RNA-Seq of Untreated Mixed Adults, digitally normalized
 - TF1_accepted_hits
 - TM_accepted_hits
 - UMA_accepted_hits
 - VAF_accepted_hits
 - VAJU_accepted_hits
 - VAM_accepted_hits
- Splice Junctions: 7

On the right, the Apollo browser window shows a track for 'EAFF_v0.5.3-Models' with a blue bar representing the gene set. The browser title is 'pollo File' and the coordinate is 0.

- Baylor Maker annotations:
 - Primary Gene Set:
 - EAFF_v0.5.3-Models
 - Other tracks that were used to generate the primary gene set
- Transcriptome/RNA-Seq
 - Transcriptome assemblies
 - Coverage plots, Mapped RNA-Seq data, Splice junctions
 - Some of the RNA-Seq libraries are stranded

Choosing reference proteins: *D. melanogaster* pepck in UniProt

UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table

BLAST Align Format Add to basket History

Protein | Phosphoenolpyruvate carboxykinase [GTP]
Gene | Pepck
Organism | *Drosophila melanogaster* (Fruit fly)
Status | Reviewed - Annotation score: ●●●○○○ - Experimental evidence at transcript levelⁱ

Annotation score is a heuristic for annotation quality

Organism-specific databases

FlyBaseⁱ FBgn0003067. Pepck.

Subcellular locationⁱ

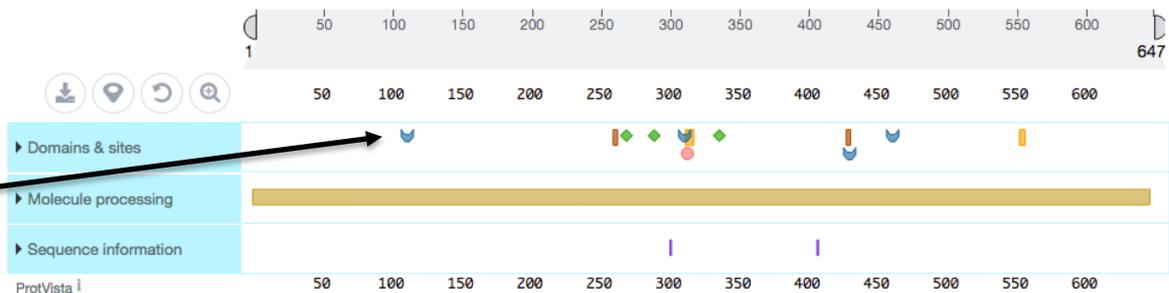
Flybase is another great resource

UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table

BLAST Align Format Add to basket History



Feature viewer gives graphical view of domains and sites

Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).

Source: <http://www.uniprot.org/uniprot/P20007>

Choosing reference proteins: *Daphnia pulex* Pepck

- GenBank record:

<https://www.ncbi.nlm.nih.gov/protein/EFX80236.1>

Lynch, M., Boore, J.L. and Grigoriev, I.V.

CONSRTM US DOE Joint Genome Institute (JGI-PGF)

TITLE Direct Submission

JOURNAL Submitted (02-FEB-2011) US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598-1698, USA

COMMENT Method: conceptual translation.

FEATURES Location/Qualifiers
source 1..652

← Treat with caution!!!

Phosphoenolpyruvate carboxykinase,

(daphnia Phosphoenolpyruvate carboxykinase)

(daphnia Phosphoenolpyruvate carboxykinase)

Manual annotation live example

BLAST dmel, dpul proteins against *E. affinis* proteins

<https://i5k.nal.usda.gov/webapp/blast/>

The screenshot displays the BLAST web interface. At the top, there are navigation links: "i5k@NAL", "Tools", "About Us", "Contact", and "Results". Below this, there are two coverage graphs: "Query Coverage Graph - EFX80236.1, BLAST Hits 1-1" and "Subject Coverage Graph - gn1|Eurytemora_affinis_protein_v0.5.3|EAFF006514-PA, BLAST Hits 1-1". Both graphs show a single hit with a score of 604 bits. Below the graphs is a table of results with the following columns: blastdb, qseqid, sseqid, pident, length, mismatch, gapopen, qstart, and qend. The table shows two entries, with the first entry highlighted in yellow. An arrow points from the text below to the sseqid cell of the first entry. To the right of the table is a "BLAST Report" section showing the query and subject sequences, along with the score and identities. An arrow points from the text below to the BLAST Report section.

blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend
euraff	EFX80236.1	EAFF006514-PA	50	612	242	10	42	65
euraff	sp P20007 PCKG_DROME	EAFF006514-PA	51.85	594	235	5	52	64

Showing 1 to 2 of 2 entries (filtered from 19 total entries)

BLAST Report

FASTA

```
>gn1|Eurytemora_affinis_protein_v0.5.3|EAFF006514-PA unnamed protein product
Length=575
194
195 Score = 604 bits (1557), Expect = 0.0, Method: Compositional matrix adjust.
196 Identities = 306/612 (50%), Positives = 401/612 (66%), Gaps = 64/612 (10%)
197
198 Query 42 KSLPSKVRFSFVEDCVKLCQPSQVHIKNGSEQENRSLIQMQQGMIESLPKMENCWLRTR 101
199 + +P VR ++ +C+P +HI +G+ +E+ +L + + + G++ LPK ENC+L RT
200 Sbjct 14 EGVPHVREWINHWADICEPQDIHIMDGTSEEDVALKKMLVRTGLIHLPKYENCFLART 73
201
```

Copy the protein 'base name'
EAFF006514 for searching in Apollo

Results are filtered by e-value; only
one protein in the *E. affinis* dataset has
a significant match

Result URL: <https://i5k.nal.usda.gov/webapp/blast/68b677fb267d4cfe93b0570dd87449f7>



Modify *E. affinis* model sequence in Apollo

- Go to Apollo URL:
<https://apollo.nal.usda.gov/euraff/jbrowse/>
 - Find mRNA of EAFF006514-PA in genome browser by pasting EAFF006514 into search box, selecting EAFF006514-RA
- Log in to Apollo
- Drag EAFF006514-RA into the yellow annotation track
- Check available evidence for model

Another approach: BLAST against the genome

<https://i5k.nal.usda.gov/webapp/blast/>

The screenshot displays the BLAST web interface. At the top, there are navigation links for 'Tools', 'About Us', 'Contact', and 'Results'. Below this, two coverage graphs are shown: 'Query Coverage Graph - EFX80236.1, BLAST Hits 1-21' and 'Subject Coverage Graph - gnl|Eurytemora_affinis|euraff_Scaff'. The main area contains a table of results and a detailed BLAST report.

blastdb	qseqid	sseqid	pident	length	mismatch	gapope
euraff	Eaff_11172013.genome_new_ids.fa	Scaffo1d33	56.41	39	17	0
euraff	sp P20007 PKG_DROME	Scaffo1d33	62.5	40	15	0
euraff	EFX80236.1	Scaffo1d33	80	30	6	0
euraff	sp P20007 PKG_DROME	Scaffo1d33	78.12	32	7	0
euraff	EFX80236.1	Scaffo1d33	44.59	74	24	2
euraff	sp P20007 PKG_DROME	Scaffo1d33	46.15	78	25	2
euraff	EFX80236.1	Scaffo1d33	38.46	26	16	0
euraff	EFX80236.1	Scaffo1d33	72.34	47	13	0

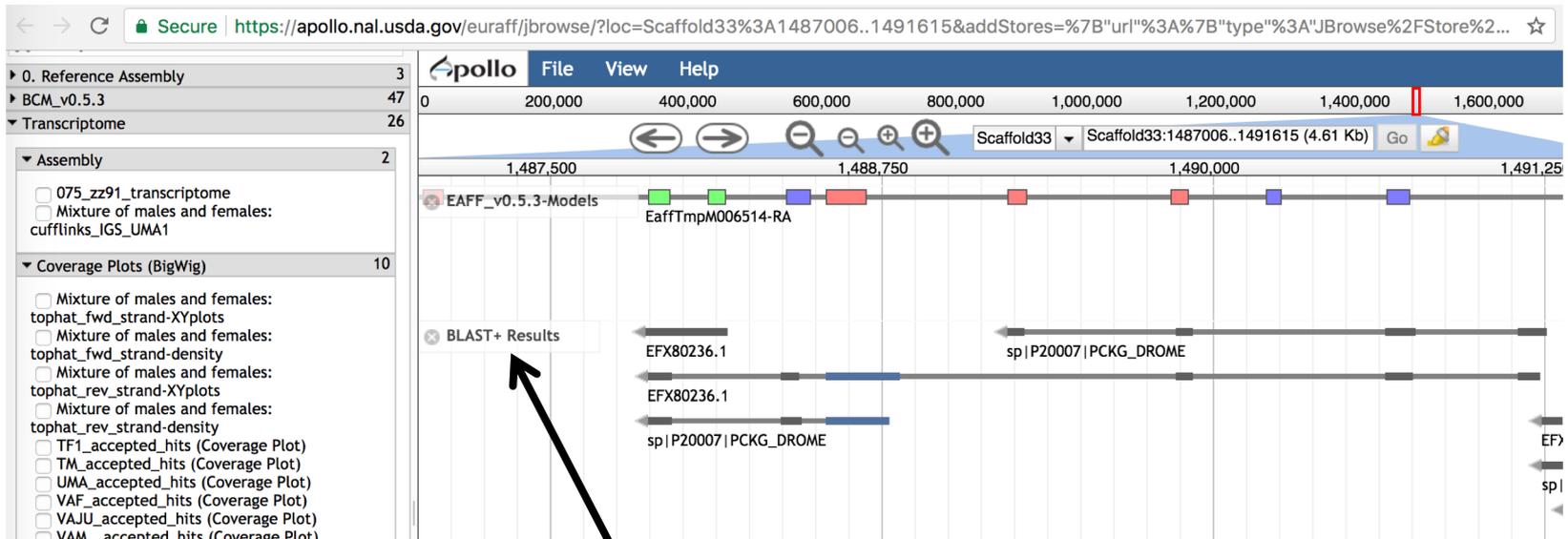
The BLAST report on the right shows details for three hits, including scores, identities, and mismatches. A black arrow points to the blue 'euraff' button in the first row of the table.

Click on blue blastdb button next to your favorite HSP to view it in JBrowse

BLAST result URL: <https://i5k.nal.usda.gov/webapp/blast/1dd580d46260410da7473f974da76a54>



Another approach: BLAST against the genome



BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

Apollo result URL: <http://tiny.cc/lwuzsy>

Create annotation in user-created annotations track

Available Tracks

- filter by text
- 0. Reference Assembly 2
- BCM_v0.5.3 47
- Transcriptome 26
- Assembly 2
 - 075_zz91_transcriptome
 - Mixture of males and females:

apollo File View Help Login

Scaffold33 Scaffold33:1481748..1500247 (18.5 Kb) Go

EAFF_v0.5.3-Models
EaffTmpM006513-RA
EaffTmpM006514-RA
EaffTmpM006515-RA

Log in with
your
Apollo
credentials

Available Tracks

- filter by text
- 0. Reference Assembly 2
- BCM_v0.5.3 47
- Transcriptome 26
- Assembly 2
 - 075_zz91_transcriptome
 - Mixture of males and females:

apollo File View Tools Help euraff_u

Scaffold33 Scaffold33:1481741..1500240 (18.5 Kb) Go

User-created Annotations

EAFF_v0.5.3-Models
EaffTmpM006513-RA
EaffTmpM006514-RA
EaffTmpM006515-RA

Drag model EaffTmpM006514-
RA to User-created Annotations
track

Modify *E. affinis* model sequence in Apollo

- Questions:
 - What evidence do you choose to check the integrity of the model?
 - Do you need additional evidence?
 - How do you evaluate whether the protein sequence is as complete as it can be?
 - Should you add/modify UTRs?

View available evidence



Model is on the reverse strand, so we can take advantage of the stranded RNA-Seq available for this species

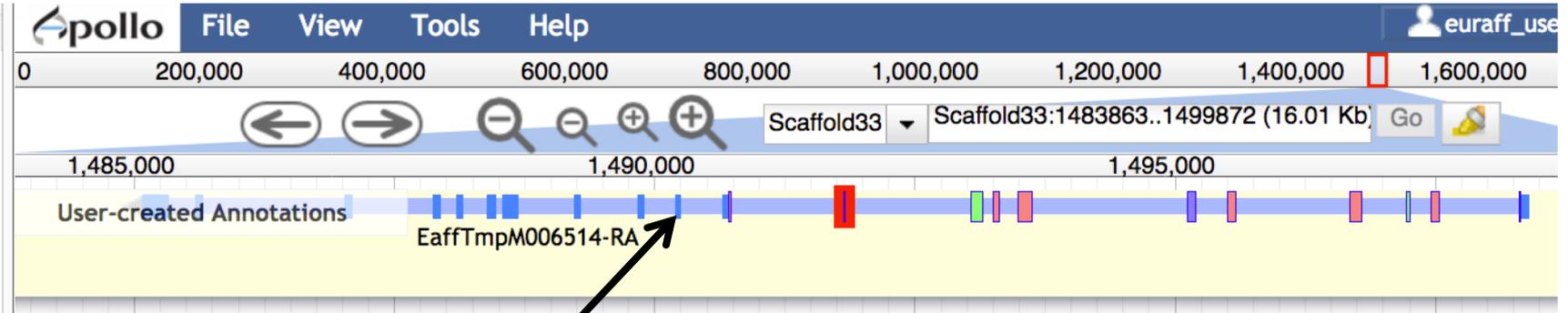
RNA-Seq and transcriptome tracks suggest that one exon is missing

Add an exon to the model

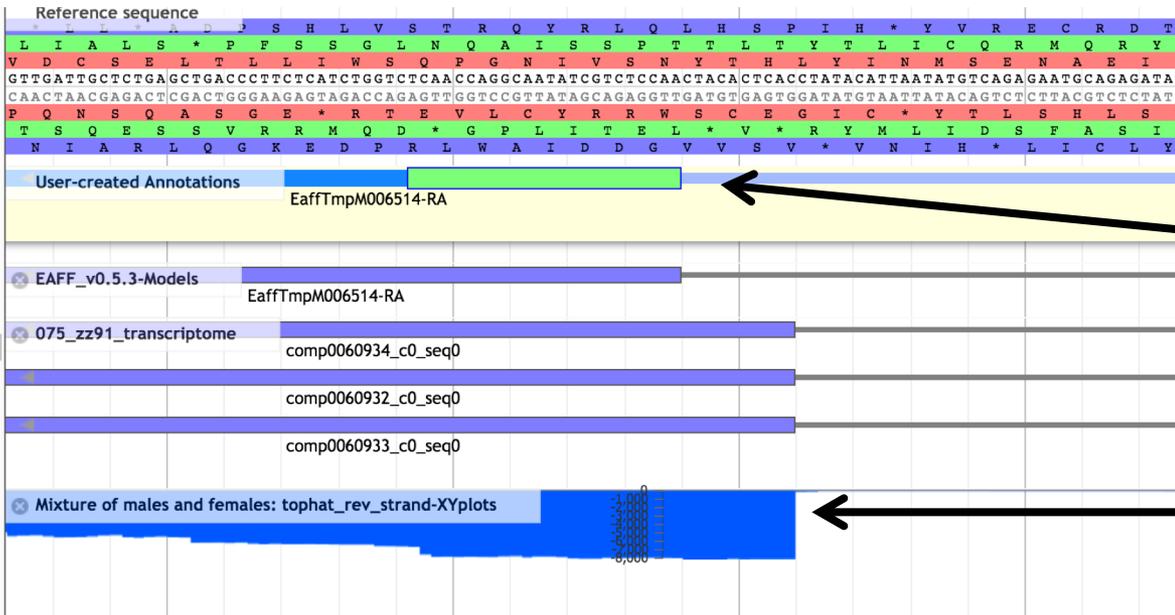
The screenshot shows the Apollo genome browser interface. The top navigation bar includes the Apollo logo, menu options (File, View, Tools, Help), and the user name 'euraff_user_admin'. The main view displays a genomic region on Scaffold33, with coordinates ranging from 0 to 1,800,000. A tooltip indicates the selected feature is an mRNA with owner 'euraff_user_admin' and last modified on 2017-08-28 at 14:34. The left sidebar lists available tracks, including Reference Assembly, Transcriptome, Assembly, Coverage Plots, and Mapped Reads. The transcriptome track shows several transcripts, with '075_zz91_transcriptome' highlighted. A red box highlights an exon in this transcriptome track. A black arrow points from this exon to a new gene model track, illustrating the process of adding an exon to the model.

Drag exon from
transcriptome track
into new gene model

Adjust exon boundary



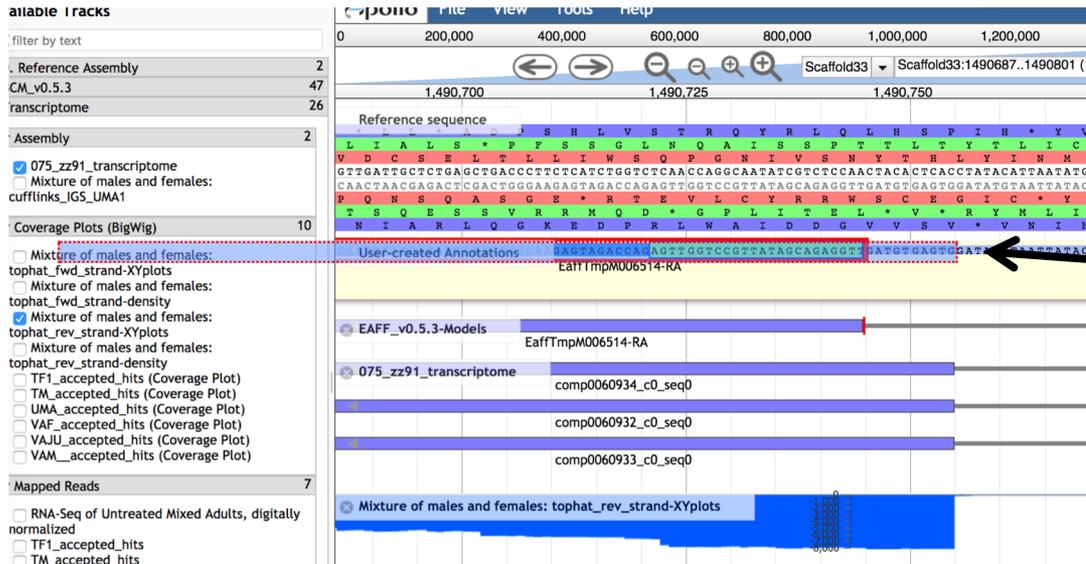
CDS sequence is now UTR –zoom in to investigate



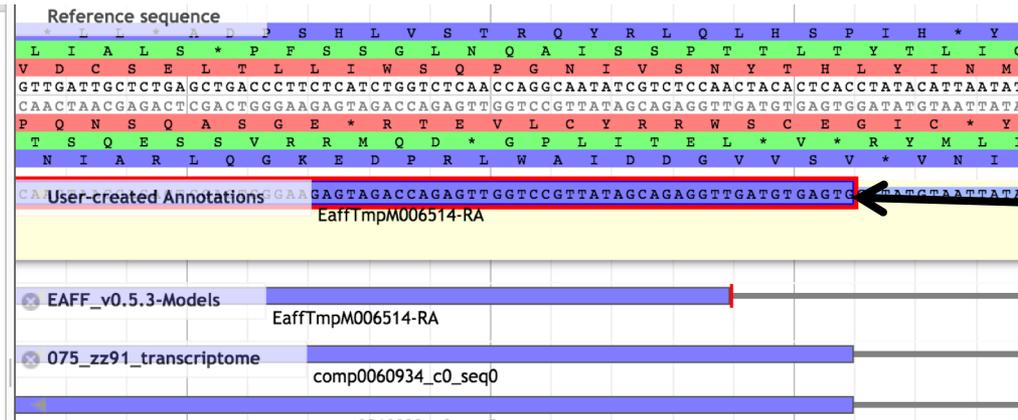
CDS frame has changed from purple to green—we need to fix this

RNA-Seq suggests we need to adjust exon boundary

Adjust exon boundary



Drag exon boundary to match RNA-Seq and transcriptome tracks



Fixed both reading frame and exon boundary

Evaluate new protein sequence

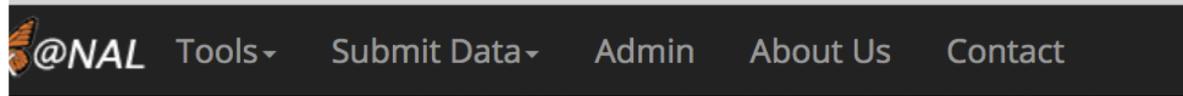
- Blast modified EAFF006514-PA sequence to NCBI's nr database
 - Make sure it doesn't match a potential contaminant
 - Get an idea whether you have the right sequence
 - Blastp home:
 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
 - Result URL:
 - <https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=DY9HRCRA015> (expires end of day 4/25)
- Once contamination is ruled out, it's better to align your sequence against a smaller set of high-quality proteins
- If you notice that parts of the protein are missing, check the 'Gaps in assembly' track in the browser

Evaluate new protein sequence

- Get *E. affinis* pepck protein sequence from old model and new model
- Align new and old sequence to dmel and dmag protein sequences
 - Clustal (<https://i5k.nal.usda.gov/webapp/clustal/>)
 - Can also use NCBI Blast
- Check alignment extent, %ID

Clustal Results

://i5k.nal.usda.gov/webapp/clustal/105850a3594e4234a21b07d93cbbd71



euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
IS-----VGDDIAWLRPDEKQQLRAI
ISGITNSQGEKKYIVAAFPSCGKTNLAMMQRLP-----VGVVGGDDIAWLRPDEKQQLRAI
ILGITDPKGEKKYITAAFPSCGKTNLAMLNPSLANYKVECVGDDIAWMKFDSSQVLRAI
ILGITNPQGGKQYIAAAPPSCGKTNLAMLTPTLPGYKVECVGDDIAWMHFDKEGRLRAI
*                               *****: : *.* ****
```

New exon added

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
NPENGGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGGVAPGTSMETNPVIA-----MNTVFKNTIFTNVASTDGGVWVEGMESSLA
NPENGGVAPGTNYATPNACYNFFLYAMLTIQKNTIFTNVAKTSDDGGVWVEGLEKEV-
*****: : * * * : : * : : * : : * : : * : : * : : * : : *
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
ERSSCIDWKGK-PWRPTSSNPAHPNSRFCTPLLNCVLDSEADPAGVPIAAILFGGRR
ERSSCIDWKGK-PWRPTSSNPAHPNSRFCTPLLNCVLDSEADPAGVPIAAILFGGRR
PNVQITDWLGK-PWTKDSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPIAAILFGGRR
TGVDITSWLGDANWTKSSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPIAAILFGGRR
. * * * * * : : * : : * : : * : : * : : * : : *
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFKQKQIMHDPFSMRPFFG-----HW
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFKQKQIMHDPFSMRPFFG-----HW
PAGVPLIYEARDWTHGVFVIGAMRSEATAAAEHKQKQIMHDPFAMRPFYGFYVAHW
PRGVPLVYEALNWKHGVFVIGASVSEATAAAEHKGRS IMHDPFAMRPFYGFYVAHW
* ****: : * * * : : * : : * : : * : : * : : * : : * : : *
```

Another exon might be missing (we're not going to handle this today)

- Clustal result URL:

<https://i5k.nal.usda.gov/webapp/clustal/49a4d63c24fd4ed3b3a67cf71a0369df>

- Scroll to bottom of page and click 'colorful' to see color-coded alignment



Using the Information Editor

- Select the model in Apollo, then right-click, and select 'Edit Information' from the drop-down menu
 - Use the 'mRNA' section
 - **Please review our naming guidelines:**
 - <https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>
 - If a naming convention exists, use it (e.g. for gene families)
 - Use name from an orthologous protein if you are sure that your gene model is orthologous.
 - Document your justification for the name in the Comments field (e.g. "88% sequence similarity via blastp to D. melanogaster pepck P20007")
 - If you create a new name, it should be unique and attributed to all orthologs (as far as possible)
 - Comments – Document what changes you performed, and your justification for the name. These notes will be visible in the OGS, so make sure that others understand them

Using the Information Editor

The screenshot shows the Apollo Information Editor interface. The browser address bar contains the URL: `euraff/browse?loc=Scaffold33%3A1482161..1498680&tracks=DNA%2CAnnotations%2Ceuraff_current_models%2Ctophat_rev_strand-XYplots&highlight=`. The browser window title is "Apollo" and the menu bar includes "File", "View", "Tools", and "Help". The main window is titled "Information Editor (alt-click)".

The "Select mRNA" dropdown menu is set to "Phosphoenolpyruvate carboxykinase". The interface is split into two columns: "gene" and "mRNA".

gene

- Name:
- Symbol:
- Description:
- Created: 2017-08-28
- Last modified: 2017-08-28

Status

Approved Delete

DBXRefs

DB	Accession
----	-----------

mRNA

- Name: Phosphoenolpyruvate carboxykinase
- Symbol: pepck
- Description:
- Created: 2017-08-28
- Last modified: 2017-08-28

Status

Approved Delete

DBXRefs

DB	Accession
----	-----------

On the left side of the browser window, a taxonomic tree is visible with categories such as "_Annelid", "_Arthrop", "_Atelocer", "_Cephala", "_Chelice", "_Cnidari", "_Craniat", "_Crustac", "_Echinoc", "_Mollusc", "_Nemato", "_Nemato", "_Onycho", "_Parazoa", "_Placozo", "_Platyhe", "_Priapul", "_Tardigr", "_Tunicat", and "_UNCATE".

Checklist for accuracy and integrity

- Check start, stop and exon boundaries (splice sites)
 - Try to fix non-canonical splice sites if possible
 - Check if you can annotate UTRs (e.g. using RNA-Seq data)
 - Check for gaps in the genome
 - If you change the genome sequence, add a justification comment to the corresponding gene model
 - Use BLAST or a multiple sequence aligner
 - To look at completeness of model
 - To verify the appropriateness of the gene name
 - In the Information editor ***mRNA*** field
 - Update the Name if appropriate
 - Add comments that describe
 - your evidence for the annotation
 - Modifications that you made to the gene model
- cf. <https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop>

[apollo-workshop](https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop)



What happens to my annotation when I'm done?

- This depends on the genome project that you're working on.
- If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
 - We are still working on this process, so if you ask us to do this, 1) it will take some time, and 2) we will probably ask you for co-authorship if you publish a paper on the OGS.
 - You can also try out the process yourself: <https://github.com/NAL-i5K/GFF3toolkit/>
 - We are working on a pipeline to submit Official Gene Sets to GenBank, where they will be archived/accessioned
- Otherwise, don't assume that your annotation will be archived.
 - If you need it to be, get in touch with us and we'll figure out what to do.
- Get in touch with us and the genome project coordinator if you're not sure about the status of a genome project.
- <https://i5k.nal.usda.gov/data-management-policy>

Thank you!

The NAL Team

- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

I5k Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

i5k Workspace@NAL advisory committee

- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!

